

15k Workspace webinar – facilitating data reuse

Monica Poelchau
National Agricultural Library
USDA-ARS
November 17th, 2020

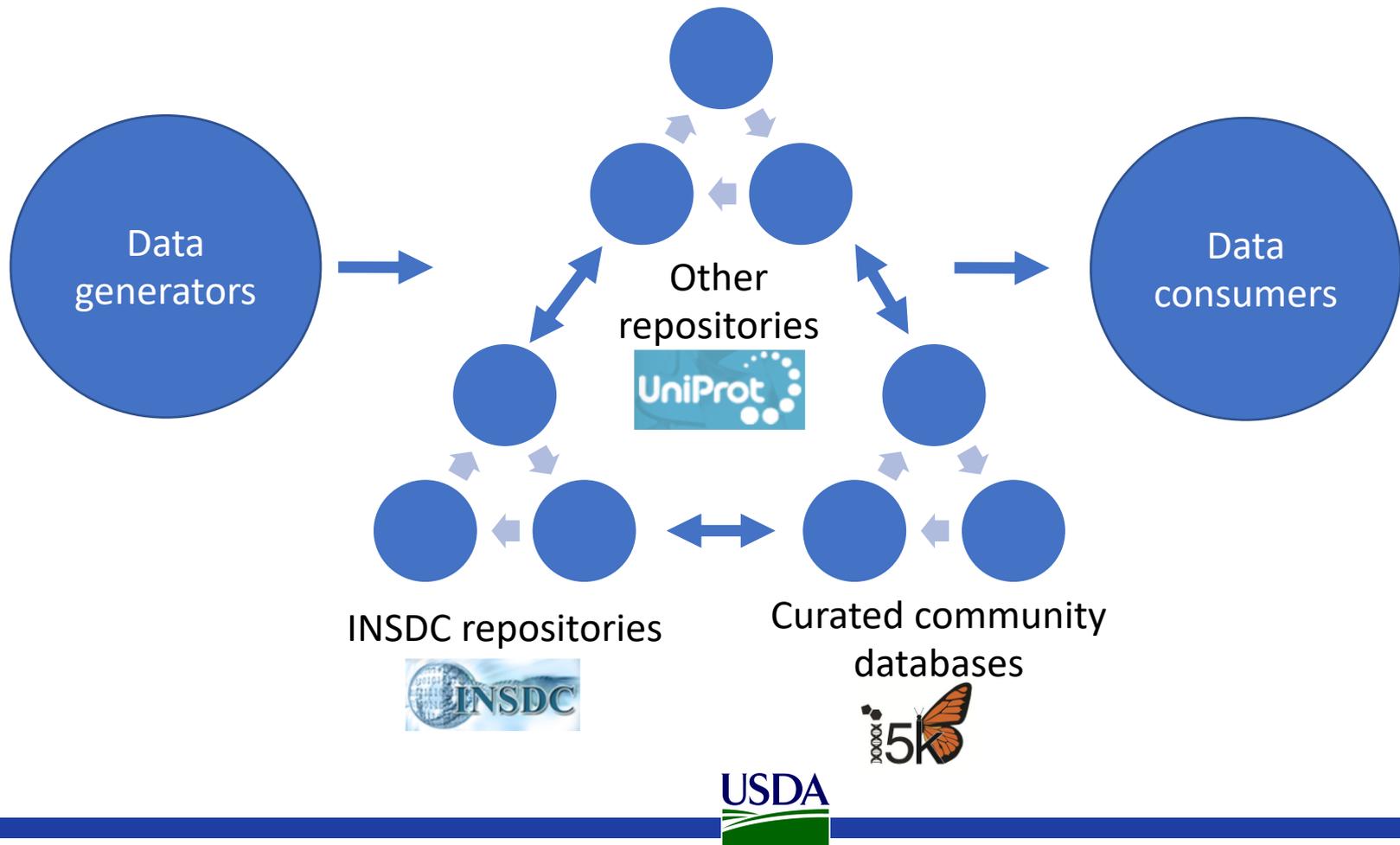


Agenda

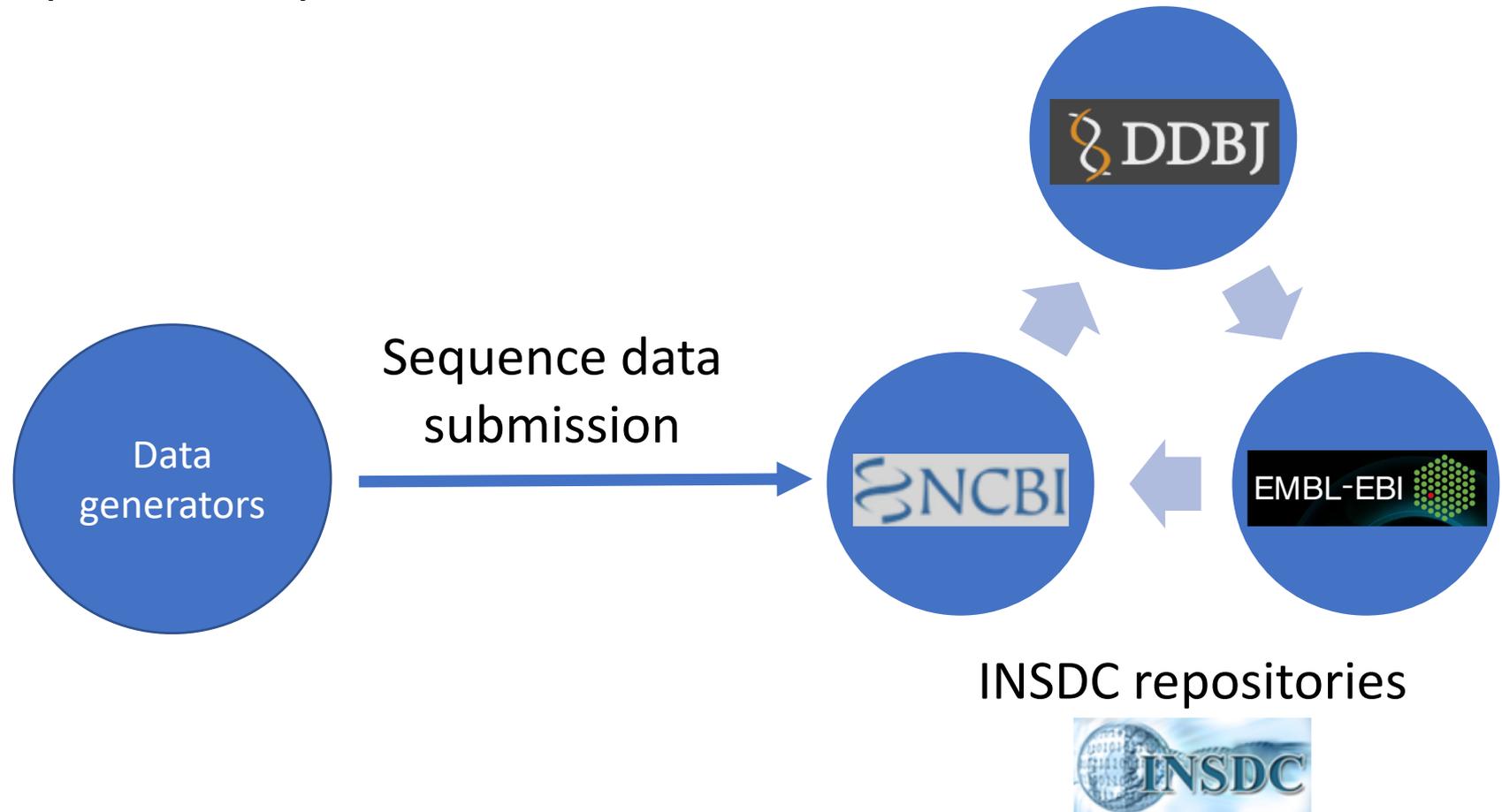
- Sharing manual annotations between databases
 - The 'data ecosystem'
 - Submission of manual annotations to NCBI's GenBank
 - Official Gene Set generation and submission
- Naming genes and proteins
 - Naming definitions
 - 15k Workspace naming guidelines

The data ecosystem

How data moves into and between databases

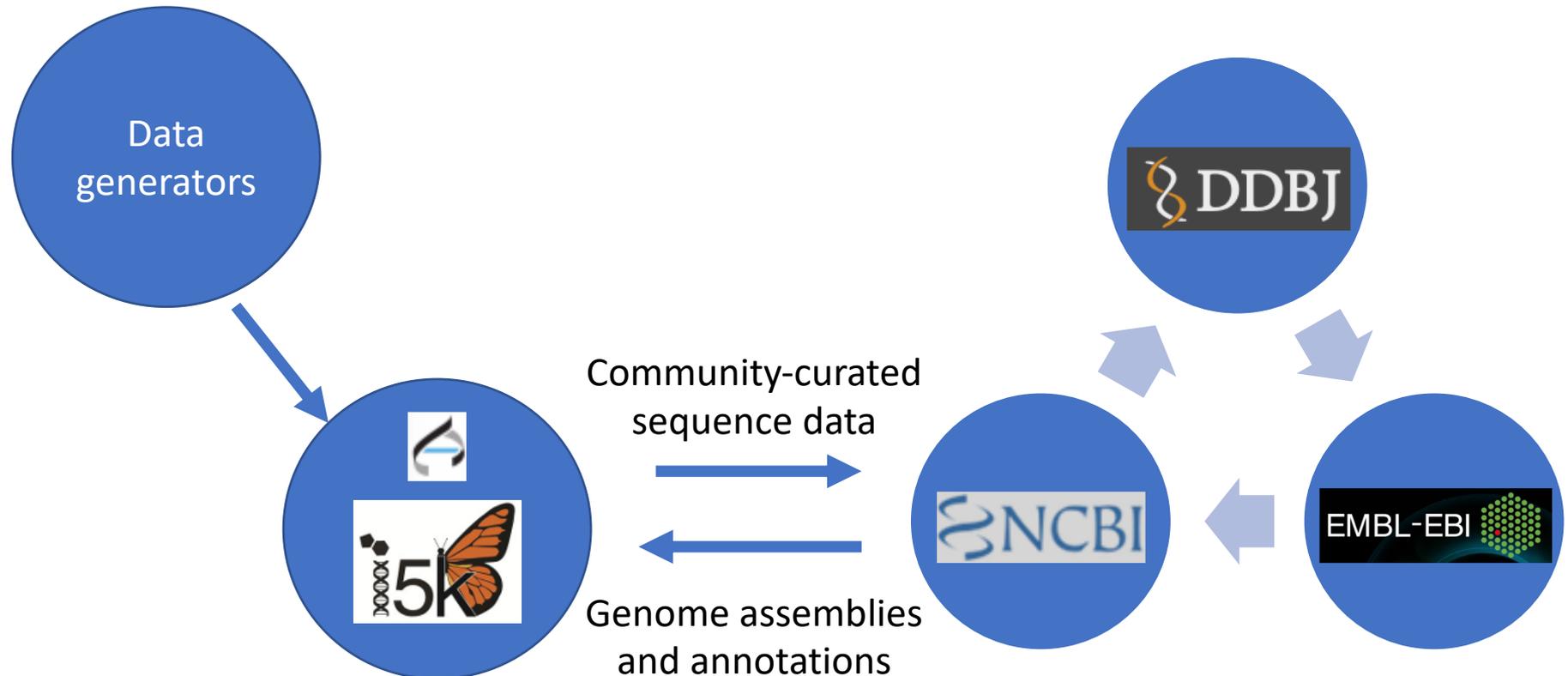


The International Nucleotide Sequence Database Consortium (INSDC)



<https://doi.org/10.1093/nar/gkx1097>

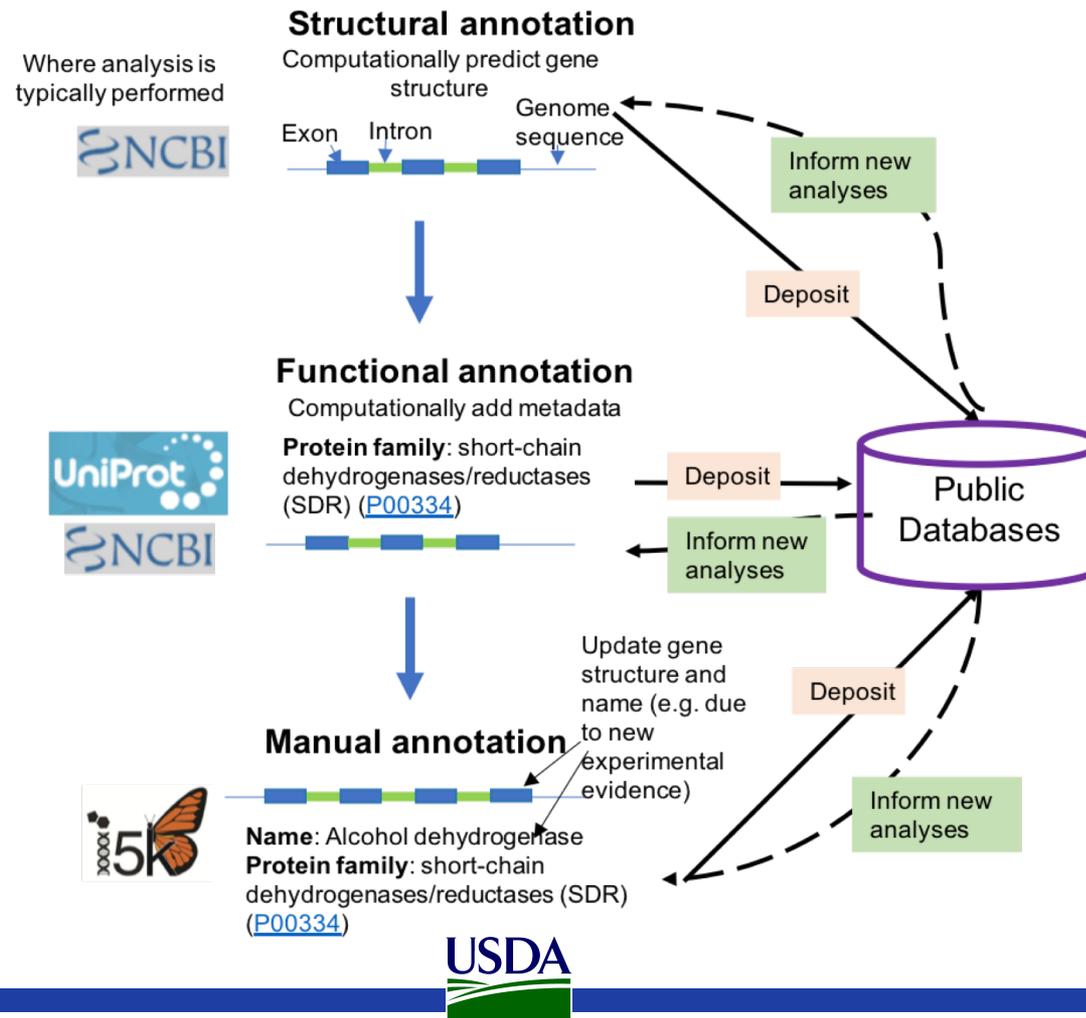
The i5k Workspace@NAL facilitates manually curated data integration



INSDC repositories



The i5k Workspace@NAL facilitates manually curated data integration



FAIR



Findable



Accessible



Interoperable

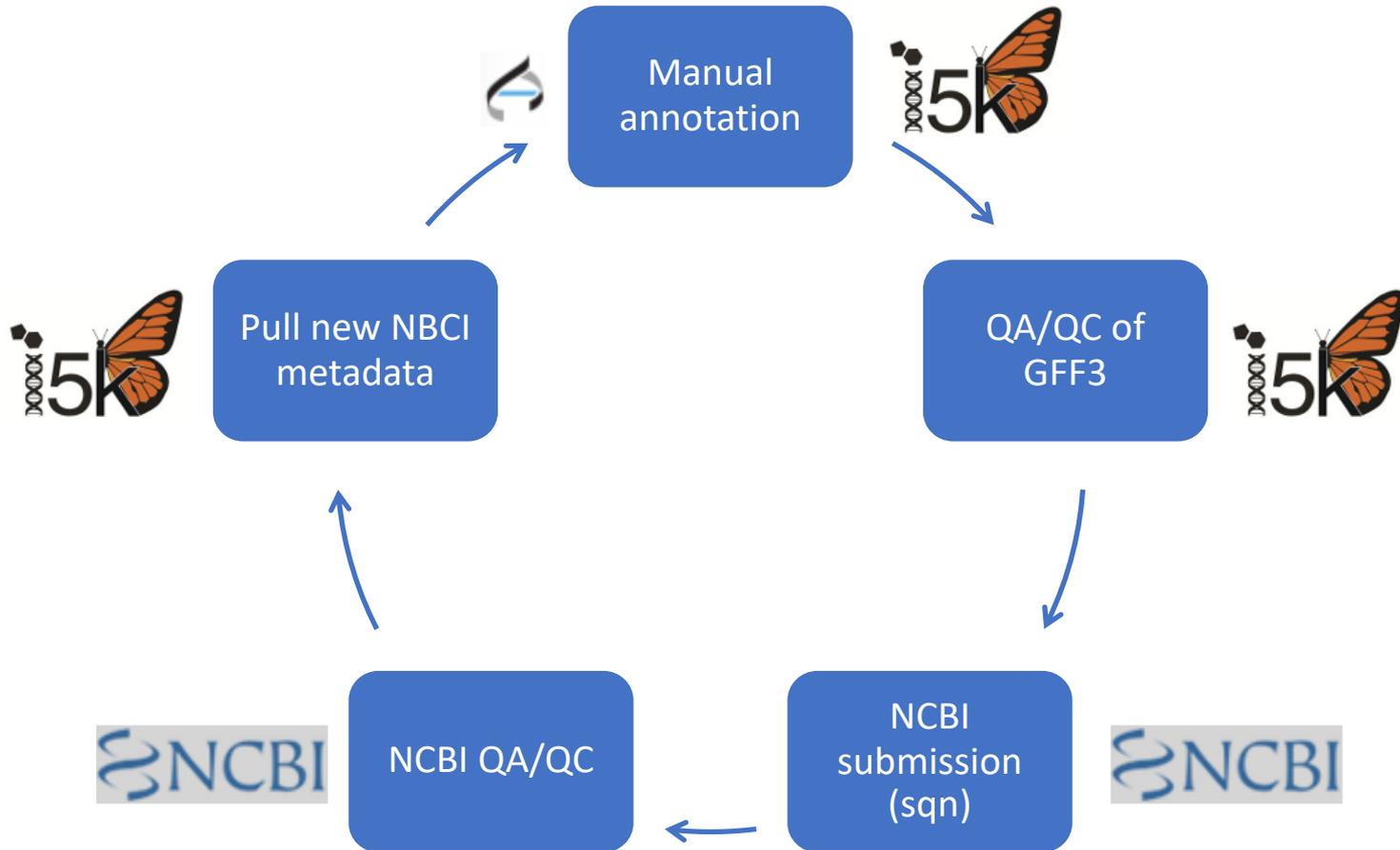


Reusable

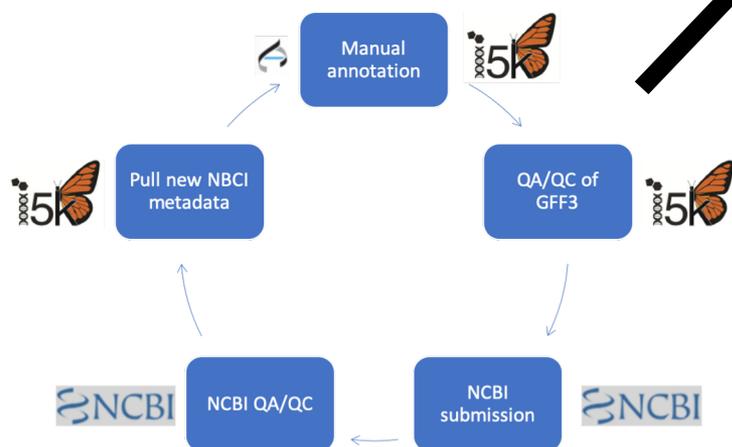
- Findable = data is human and machine readable and attached to persistent identifiers
- Accessible = data can be found and retrieved by humans and machines using standard formats
- Interoperable = data can be exchanged and used between systems
- Reusable = data can be used by others

Data integration between the i5k Workspace and NCBI's GenBank

Manual annotation QA/QC and submission



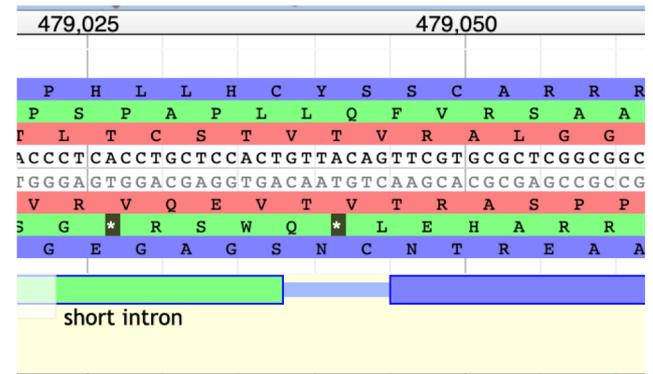
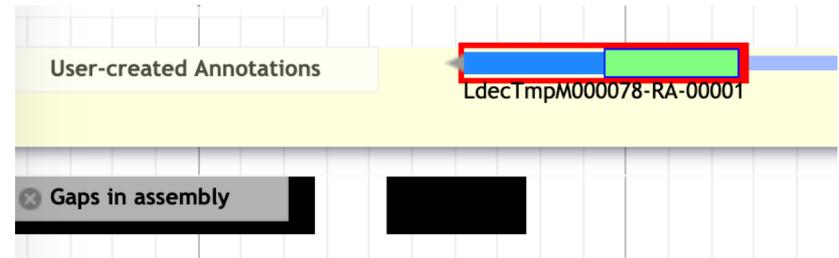
Typical QA/QC process



- Programs for general structural QC and fixes:
 - GFF3toolkit – frequent errors found in Apollo
 - table2asn_GFF – general issues, and NCBI-specific
- Program for QC of names and symbols:
 - table2asn_GFF
- NCBI-specific metadata
 - Custom scripts
- This process can be time-consuming!

Typical issues in manual annotations

- Feature begins or ends in gap
- Introns < 10 bp
- Duplicate transcripts
- Pseudogene markup
- ***Gene/protein names do not follow NCBI guidelines***
- Notes or descriptions need to be discarded



Result of a successful GenBank submission

← → ↻ ncbi.nlm.nih.gov/protein/RLZ02283.1

Odorant receptor 58 [Cephus cinctus]

GenBank: RLZ02283.1
[Identical Proteins](#) [FASTA](#) [Graphics](#)

Go to:

LOCUS RLZ02283 411 aa linear INV 24-OCT-2018
DEFINITION Odorant receptor 58 [Cephus cinctus].
ACCESSION RLZ02283
VERSION RLZ02283.1
DBLINK BioProject: [PRJNA168335](#)
BioSample: [SAMN02905554](#)
DBSOURCE accession [KB467292.1](#)
KEYWORDS .
SOURCE Cephus cinctus (wheat stem sawfly)
ORGANISM [Cephus cinctus](#)
Eukaryota; Metazoa; Ecdysozoa; Arthropoda; Hexapoda; Insecta;
Pterygota; Neoptera; Holometabola; Hymenoptera; Cephoidae;
Cephidae; Cephus.
REFERENCE 1 (residues 1 to 411)
AUTHORS Robertson,H.M., Robinson,G.E., Wanner,K.W. and Walden,K.K.O.
TITLE The Genome of the Wheatstem Sawfly, Cephus cinctus
JOURNAL Unpublished
REFERENCE 2 (residues 1 to 411)
AUTHORS Robertson,H.M., Robinson,G.E., Wanner,K.W. and Walden,K.K.O.
TITLE Direct Submission
JOURNAL Submitted (31-AUG-2012) Entomology, University of Illinois at
<https://www.ncbi.nlm.nih.gov/protein/RLZ02282.1>



Result of a successful GenBank submission – i5k Workspace page

CCIN027589, CCIN027589 (gene) *Cephus cinctus*

[Overview](#)

[Sequences](#)

[Transcripts](#)

Transcripts

The following features are part of this gene:

CCIN027589-RA

Details

Name Odorant receptor 58

ID CCIN027589-RA

Type mRNA

Dbxref NCBI_GP:RLZ02283.1

Analysis [Cephus cinctus annotations cepcin_OGSv1.1](#)
Source: [Cephus cinctus genome assembly Ccin1 \(GCF_000341935.1\)](#)

Annotator Comments Note: manually curated model, revised mRNA compared to XM_015753704.2; manually curated model, revised mRNA compared to XM_015753705.2

owner hrobertson

<https://i5k.nal.usda.gov/CCIN027556>



Manual annotations accepted so far

- *Cephus cinctus*
- *Diachasma alloeum*
- *Ephemera danica*
- *Frankliniella occidentalis*
- *Halyomorpha halys*
- *Hyalella azteca*
- *Laodelphax striatella*
- *Oncopeltus fasciatus*

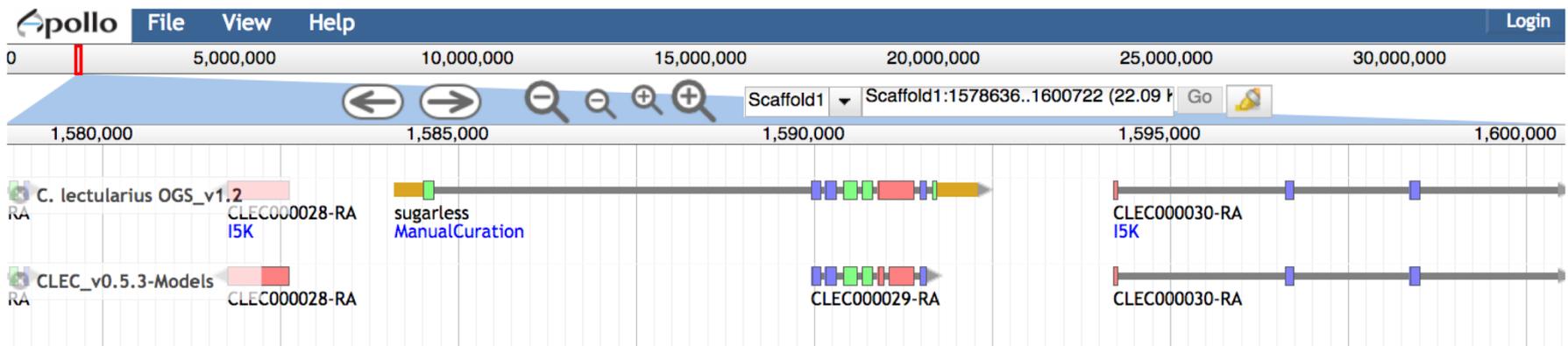
Resources

- Programs:
 - ftp://ftp.ncbi.nih.gov/toolbox/ncbi_tools/converters/by_program/table2asn_GFF/
 - <https://github.com/NAL-i5K/GFF3toolkit>
- Submitting GFF3 files to NCBI:
 - https://www.ncbi.nlm.nih.gov/sites/genbank/genomes_gff/
- GenBank submission template form:
 - <https://submit.ncbi.nlm.nih.gov/genbank/template/submission/>
- NCBI Genome submission portal:
 - <https://submit.ncbi.nlm.nih.gov/subs/genome/>

Official Gene Set generation

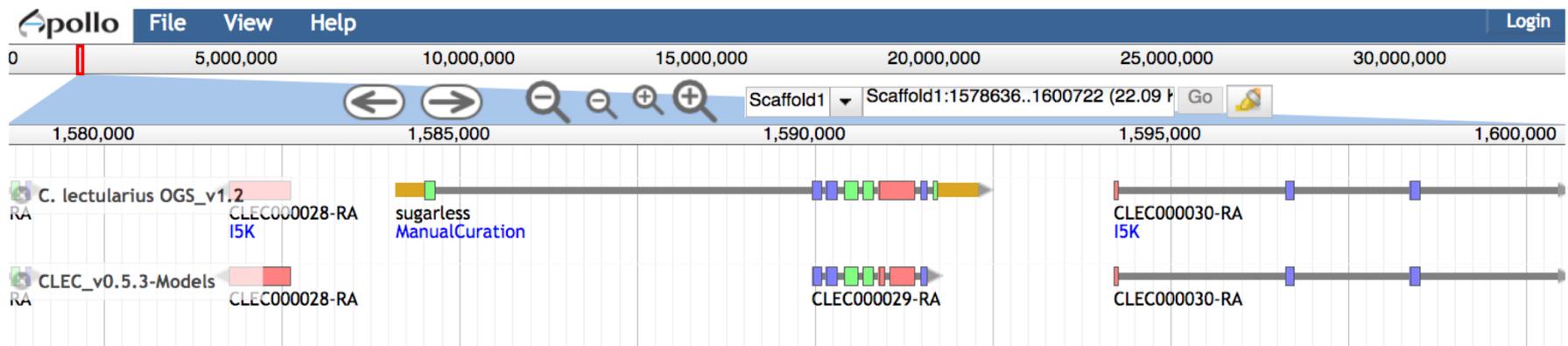
The Official Gene Set (OGS) – what is it?

- Loose definition: The best known representation of the set of gene models for a given genome assembly
- When the i5k Workspace generates an OGS, this is a merge between one gene set (usually computationally predicted), and a set of manually validated annotations (usually from the Apollo software)

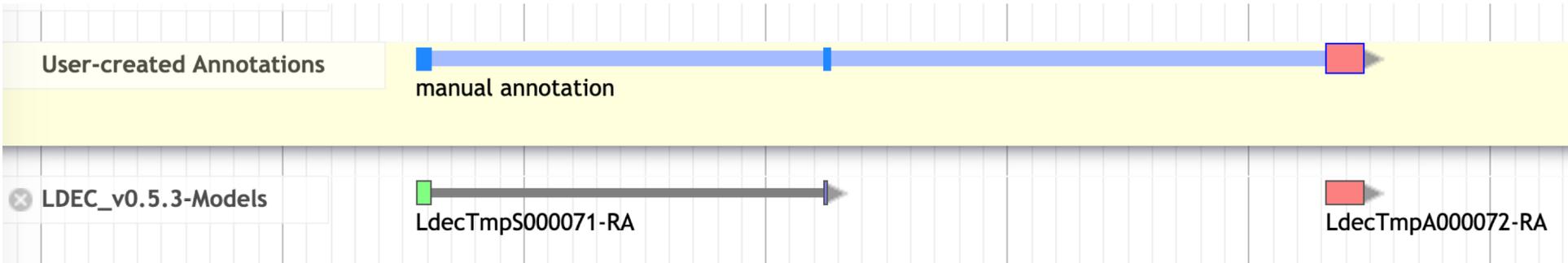


Why generate an Official Gene Set?

- This depends on your genome community's needs.
- If several groups want to perform downstream analyses, it helps to have an authoritative 'reference gene set' for your community, rather than multiple competing gene sets



Our OGS generation process – the GFF3toolkit

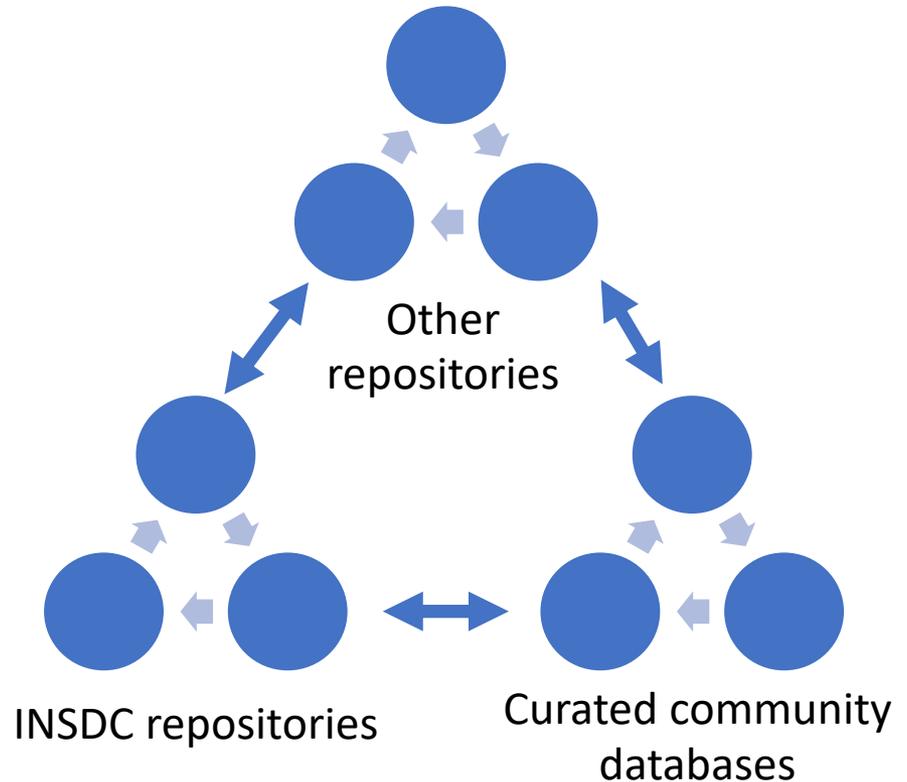


1. Check for coordinate overlap AND sequence similarity between manual annotations and reference annotations
2. If there is both, the manual model replaces the reference annotation(s)
3. Changes between the 'reference' annotation and the merged gene set are categorized into 'simple replacement', 'merge replacement', 'split replacement', 'add', and 'multi-isoform' replacement

<https://github.com/NAL-i5K/GFF3toolkit> (Mei-Ju Chen, Li-Mei Chiang)

OGS preservation and archiving

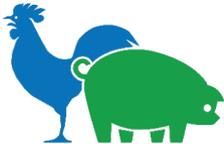
- We host the OGS at the i5k Workspace@NAL
- For preservation and archiving of the nucleotide and protein sequences, we submit the OGS or the manual annotations to NCBI
- For preservation and cataloging of the whole dataset, we submit the OGS to the Ag Data Commons



The Ag Data Commons...

- Is a catalog and data repository for USDA-funded research data
- Provides expert services for creating, curating, and enabling access to complete and machine-readable scientific metadata (FAIR data)
- Creates infrastructure for linking information, data, publications, people,...
- Helps the USDA-funded research community meet public access requirements
- Provides a DOI for data submissions

<https://data.nal.usda.gov/>



Gene and protein naming guidelines

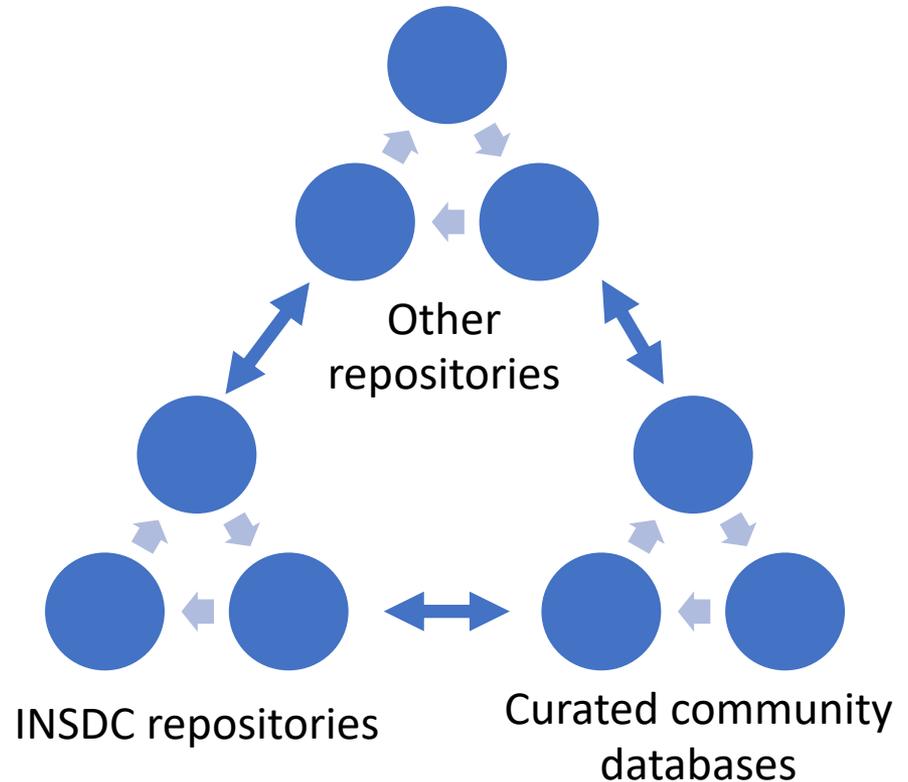
Naming standards

- Several larger genome communities have committees (sometimes funded) for naming standard development and enforcement
 - E.g. in human, vertebrates, fly, maize
- 15k Workspace doesn't have such a committee.
 - Your name gatekeepers are mainly NCBI and myself
- We have adapted the International Protein Nomenclature Guidelines for Apollo use:
 - <https://i5k.nal.usda.gov/i5k-workspace-gene-and-protein-naming-guidelines>



Naming standards – why?

- Name carries important information about protein or gene function
- Name will often be propagated to other species – needs to make sense in their context, as well
- Helps to improve consistency across taxa/databases

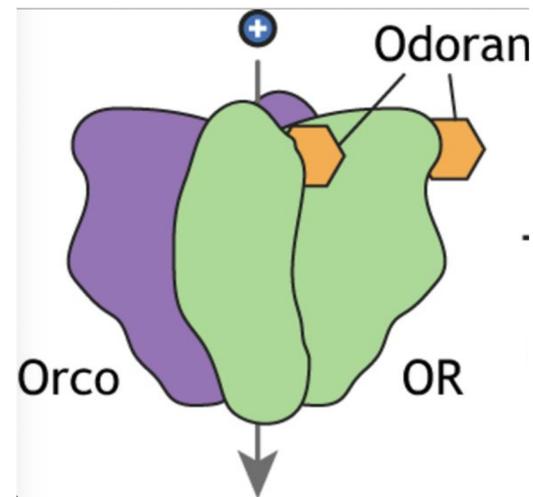


Definitions

Gene and protein names

- Provides a brief description of a gene or protein.
- Names can be applied to both genes and proteins.
- Ideally is unique, unambiguous, and can be attributed to orthologs from other species
- Should not describe a phenotype, anatomical features, or taxon-specific characteristics.

- Example gene name: *Odorant receptor coreceptor*
- Example protein name: *Odorant receptor coreceptor*



https://www.ncbi.nlm.nih.gov/genome/doc/internatprot_nomenguide/

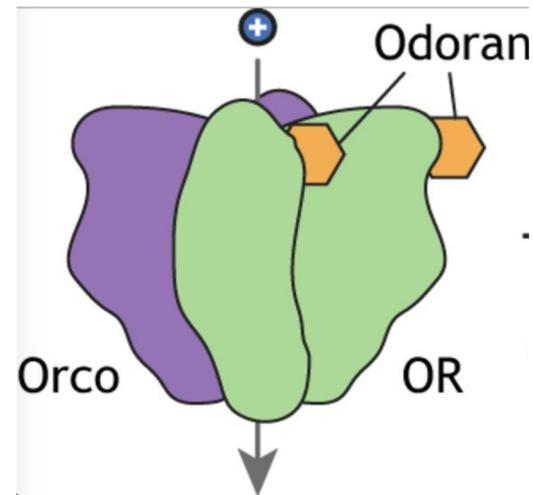
https://jeb.biologists.org/content/223/Suppl_1/jeb208215.figures-only

Gene symbols

- A gene symbol is a short form of the gene or protein name.
- In eukaryotes, symbols typically apply only to genes.
- Example gene name:
Odorant receptor coreceptor
- Example protein name:
Odorant receptor coreceptor

https://www.ncbi.nlm.nih.gov/genome/doc/internatprot_nomenguide/

- Example gene symbol:
Orco



https://jeb.biologists.org/content/223/Suppl_1/jeb208215.figures-only

Accessions

- Accession: A local identifier.
 - For example, XP_015127536.1 is an accession that refers to a specific entry in NCBI's protein database – but it could refer to something different in an unrelated database.
 - The full URI (unique resource identifier) for this accession begins with a URI pattern:
https://www.ncbi.nlm.nih.gov/protein/XP_015127536.1

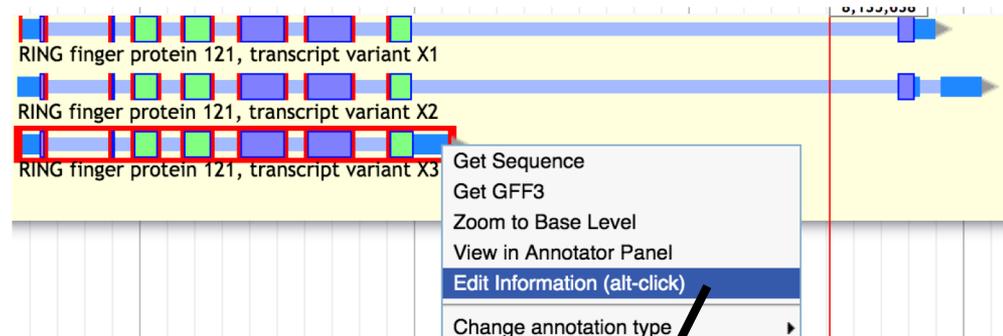
- Sometimes, Apollo will propagate an accession number to the 'Name' field. You do not need to maintain this.

Reference: <https://doi.org/10.1371/journal.pbio.2001414>

15k Workspace Guidelines

Gene and protein names

- For i5k Workspace annotation in Apollo:
 - Open the information editor for the gene you're editing
 - Enter the protein name under 'Name' in the mRNA panel
 - Enter the gene name under 'Name' in the gene panel



The "Information Editor" window is divided into two panels: "gene" and "mRNA". The "Select mRNA" dropdown at the top is set to "RING finger protein 121, transcript variant X2".

gene		mRNA	
Name	<input type="text"/>	Name	RING finger protein 121, transcrip
Symbol	<input type="text"/>	Symbol	<input type="text"/>
Description	<input type="text"/>	Description	<input type="text"/>
Created	2019-06-12	Created	2019-06-12
Last modified	2019-06-12	Last modified	2019-06-12
Status		Status	
<input type="radio"/> Approved <input type="radio"/> Delete		<input type="radio"/> Approved <input type="radio"/> Delete	
DBXRefs		DBXRefs	
DB	Accession	DB	Accession

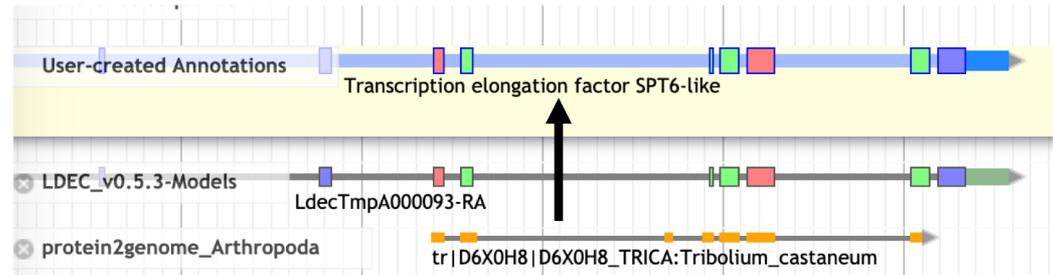
15k Workspace Guidelines – Naming use cases

1. Adopting a name from an ortholog
2. Multi-isoform genes
3. Fragmented genes
4. Coining new names
5. Gene families

I5k Workspace Guidelines - Names

Are you adopting a name from an ortholog?

- You can re-use existing, established names (e.g. from *Tribolium castaneum*)
- Don't add a species prefix (although okay to use in your manuscript for clarity)
- If you want to imply uncertainty, you can append '-like' to the name



- Good: Transcription elongation factor SPT6 🌱
- Okay: Transcription elongation factor SPT6-like
- Avoid: “Ldec-transcription elongation factor SPT6” or “similar to transcription elongation factor SPT6” 🚫

<https://i5k.nal.usda.gov/i5k-workspace-gene-and-protein-naming-guidelines>

15k Workspace Guidelines - Names

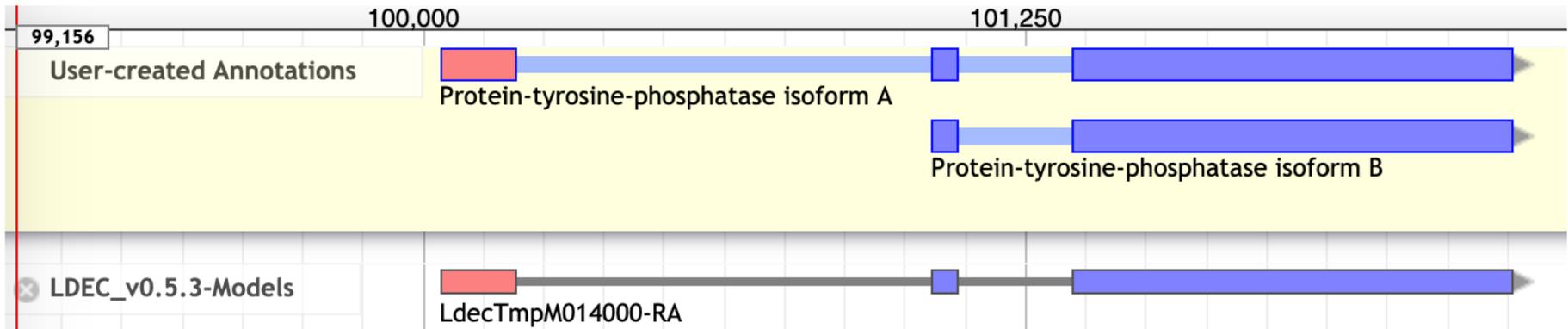
- **Are you naming a gene with multiple isoforms?**

- use the suffix “isoform A”, “isoform B”, etc.

- Good: Protein-tyrosine-phosphatase isoform A



- Avoid: Protein-tyrosine-phosphatase RB

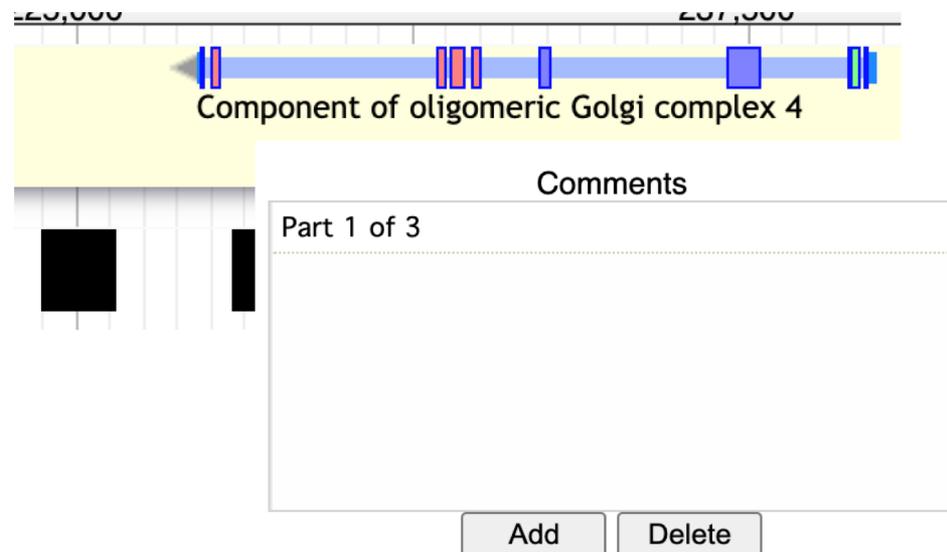


<https://i5k.nal.usda.gov/i5k-workspace-gene-and-protein-naming-guidelines>

15k Workspace Guidelines - Names

- **Are you naming a fragmented gene?**
 - include a comment 'Part X of Y', where Y is the total number of fragments, and X is the ordinal number for that gene.
 - Don't add 'partial' or 'part of' to the name.

- Good: Glycerate kinase 🟢
- Avoid: Glycerate kinase, partial 🚫

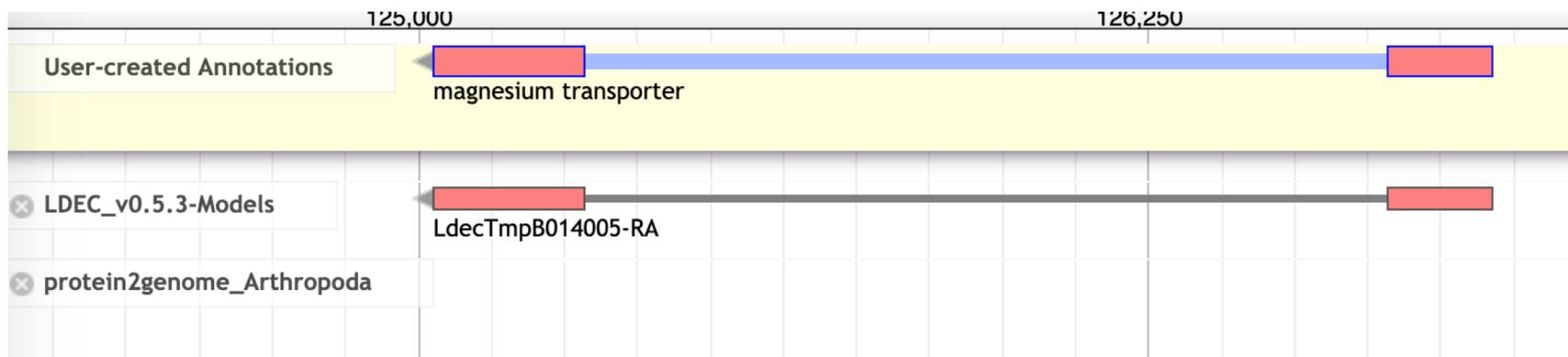


The screenshot shows a genomic track with a yellow background. A blue bar represents a gene, labeled "Component of oligomeric Golgi complex 4". The gene is fragmented into several segments, indicated by vertical bars. A comment box is open, titled "Comments", and contains the text "Part 1 of 3". Below the comment box are two buttons: "Add" and "Delete".

<https://i5k.nal.usda.gov/i5k-workspace-gene-and-protein-naming-guidelines>

15k Workspace Guidelines - Names

- **Are you creating a new name?**
- Only if there is no existing name yet in an ortholog
- Choose a name that could be propagated to all orthologous proteins; try not to make it species- or tissue-specific
- **Good: “magnesium transporter”** 
- **Avoid: “diapause-associated protein”** 



<https://i5k.nal.usda.gov/i5k-workspace-gene-and-protein-naming-guidelines>

15k Workspace Guidelines - Names

- **Are you naming a gene from a gene family?**

- Check if a naming system already exists:
<http://www.uniprot.org/docs/nomlist.txt>
- Use Arabic numbers to specify the different members encoded by a multigene family.

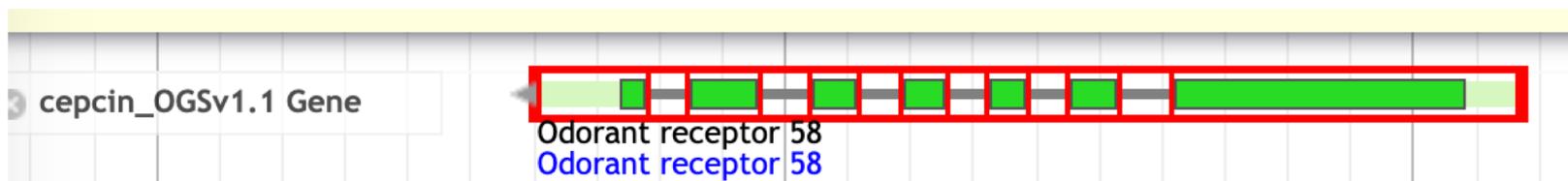
- Good:

- Odorant receptor 58
- Odorant receptor 59



- Avoid:

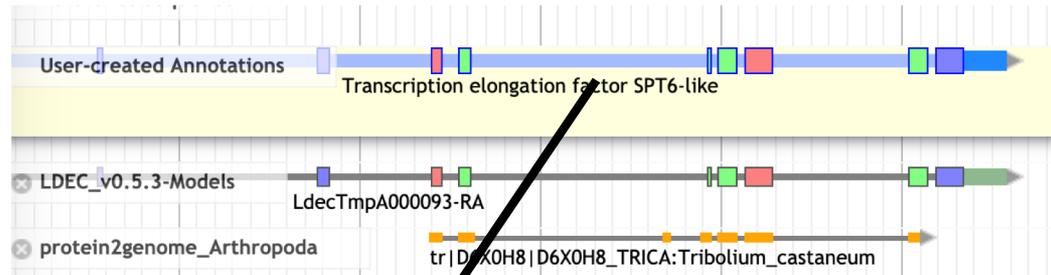
- Odorant receptor IV



<https://i5k.nal.usda.gov/i5k-workspace-gene-and-protein-naming-guidelines>

Symbols

- For i5k Workspace annotation in Apollo:
 - Open the information editor for the gene you're editing
 - Enter the symbol either under the gene or mRNA panel



The 'Information Editor' window is divided into two panels: 'gene' and 'mRNA'. The 'Select mRNA' dropdown is set to 'Transcription elongation factor SPT6-like'. The 'gene' panel has the following fields: Name (Transcription elongation factor SP), Symbol (Spt6), Description (empty), Created (2020-11-17), and Last modified (2020-11-17). The 'mRNA' panel has: Name (Transcription elongation factor SP), Symbol (empty), Description (empty), Created (2020-11-17), and Last modified (2020-11-17). Both panels have a 'Status' section with radio buttons for 'Approved' and 'Delete'.

gene		mRNA	
Name	Transcription elongation factor SP	Name	Transcription elongation factor SP
Symbol	Spt6	Symbol	
Description		Description	
Created	2020-11-17	Created	2020-11-17
Last modified	2020-11-17	Last modified	2020-11-17
Status		Status	
<input type="radio"/> Approved <input type="radio"/> Delete		<input type="radio"/> Approved <input type="radio"/> Delete	

15k Workspace Guidelines - Symbols

- Are abbreviations of the gene or protein name.
 - We do not recommend coining new symbols for newly named genes.
 - However, if a name from an orthologous gene was adopted, you may use this gene's symbol, as well.
- Good: Pepck, Ser12 
 - Avoid: Clec-Pepck 

<https://i5k.nal.usda.gov/i5k-workspace-gene-and-protein-naming-guidelines>

Other naming resources

- 15k Workspace: <https://i5k.nal.usda.gov/i5k-workspace-gene-and-protein-naming-guidelines>
- AphidBase: <https://bipaa.genouest.org/is/how-to-annotate-a-genome/>
- VectorBase: <https://www.vectorbase.org/content/gene-metadata-form>
- HGD: <http://hymenopteragenome.org/>
- FlyBase: <https://wiki.flybase.org/wiki/FlyBase:Nomenclature>
- NCBI: https://www.ncbi.nlm.nih.gov/genome/doc/international_nomenclature_guide/

Thank you!

- AgBioData (<https://www.agbiodata.org/>)
- The NAL Team
- i5k Coordinating Committee
- I5k Workspace working group
- Apollo & JBrowse Development Teams
- GMOD/Tripal community
- All of our users and contributors!

Contact us:

<https://i5k.nal.usda.gov/contact>

i5k@ars.usda.gov

